BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 2: e27087
doi: 10.3897/biss.2.27087

OPEN ACCESS

Conference Abstract

Options to Apply the IGSN Model to Biodiversity Data

Donald Hobern‡, Andrea Hahn‡, Tim Robertson‡

‡ Global Biodiversity Information Facility, Copenhagen, Denmark

Corresponding author: Donald Hobern (dhobern@gbif.org)

Received: 30 May 2018 | Published: 31 May 2018

Citation: Hobern D, Hahn A, Robertson T (2018) Options to Apply the IGSN Model to Biodiversity Data. Biodiversity Information Science and Standards 2: e27087. https://doi.org/10.3897/biss.2.27087

Abstract

For more than a decade, the biodiversity informatics community has recognised the importance of stable resolvable identifiers to enable unambiguous references to data objects and the associated concepts and entities, including museum/herbarium specimens and, more broadly, all records serving as evidence of species occurrence in time and space. Early efforts built on the Darwin Core *institutionCode*, *collectionCode* and *catalogueNumber* terms, treated as a triple and expected to uniquely to identify a specimen. Following review of current technologies for globally unique identifiers, TDWG adopted Life Science Identifiers (LSIDs) (Pereira et al. 2009). Unfortunately, the key stakeholders in the LSID consortium soon withdrew support for the technology, leaving TDWG committed to a moribund technology. Subsequently, publishers of biodiversity data have adopted a range of technologies to provide unique identifiers, including (among others) HTTP Universal Resource Identifiers (URIs), Universal Unique Identifiers (UUIDs), Archival Resource Keys (ARKs), and Handles. Each of these technologies has merit but they do not provide consistent guarantees of persistence or resolvability. More importantly, the heterogeneity of these solutions hampers delivery of services that can treat all of these data objects as part of a consistent linked-open-data domain.

The geoscience community has established the System for Earth Sample Registration (SESAR) that enables collections to publish standard metadata records for their samples and for each of these to be associated with an International Geo Sample Number (IGSN http://www.geosamples.org/igsnabout). IGSNs follow a standard format, distribute responsibility for uniqueness between SESAR and the publishing collections, and support resolution via HTTP URI or Handles. Each IGSN resolves to a standard metadata page, roughly equivalent in detail to a Darwin Core specimen record. The standardisation of identifiers has allowed the community to secure support from some journal publishers for promotion and use of IGSNs within articles.

The biodiversity informatics community encompasses a much larger number of publishers and greater pre-existing variation in identifier formats. Nevertheless, it would be possible to deliver a shared global identifier scheme with the same features as IGSNs by building off the aggregation services offered by the Global Biodiversity Information Facility (GBIF). The GBIF data index includes normalised Darwin Core metadata for all data records from registered data sources and could serve as a platform for resolution of HTTP URIs and/or Handles for all specimens and for all occurrence records. The most significant trade-off requiring consideration would be between autonomy for collections and other publishers in how they format identifiers within their own data and the benefits that may arise from greater consistency and predictability in the form of resolvable identifiers.

Keywords

object identifiers, IGSN, GBIF data index

Presenting author

Donald Hobern

References

- Pereira R, Richards K, Hobern D, Hyam R, Belbin L, Blum S (2009) TDWG Life Sciences Identifiers (LSID) Applicability Statement. https://github.com/tdwg/guid-as/blob/master/lsid/applicability_statement.doc. Accessed on: 2018-4-03.